Spam in UGC based social networks

- Spam in UGC (User Generated Content) based online social networks (OSN), such as blogs, has increased explosively on the Internet in recent years.
  - Spam repeatedly posted to a blog system to pollute top article list and the most recent article list
  - Spam using popular keywords to rank high in search results
  - Estimated that 75% active posts in Google BlogSpot are spam

- Spammers in UGC social networks is easy to conduct but hard to prevent comparing with email spam:
  - Easier to collect spamming targets by identifying UGC sites
  - Easier to post a spam article than to send a spam email
  - No systematic prevention like email server authentication
  - CAPTCHA is only for user registration instead of posting
  - Lacks legal restrictions on spamming in UGC sites

Problem Statement

- Content-based spam detection limitations for blog spam filtering:
  - New labeling data is not constantly available as in email system
  - Spam copied from recent web articles are hard to be classified

- Research Problem:
  - Understanding the inherent patterns of blog spamming behavior
  - Effectively detect blog spam in adapting to Internet environment

Blog Measurement & Analysis

- Analyze the database dump of Yahoo! Hong Kong (HK) Blog
  - a large blog site in Asia
  - over 6 million posts and nearly 400 thousand users (Table 1)
  - 325 days (from October 2008 to August 2009)

- Constant posting pattern of spammers (except for midnight) (Fig. 1)
- Spammers are paid for posting
- Spam consumes up to 84.2% of posting upload bandwidth (Fig. 2)
- Overwhelming resource consumption of spammers
- Spammers post more frequently than other users, with constant frequency (Fig. 4)

- Using bots-like automatic tools
- Most of the posts by spammers contain links to customers (Fig. 6)
- Avoid text-only posts that cannot get any clicks

Blog Spam Detection

- Table 3 shows five sets of features summarized based on measurement analysis for spam classification

Spam Detection Results

- Labeled 2167 users with links in posts
- All experiments are performed by using 10-fold cross-validation to avoid biased selections of training and testing sets

- Table 4 shows the classification results by using all the features:
  - Decision Tree has the highest precision rate of 97.5% (with 98.9% recall)
  - LR has the highest recall rate of 99.5% (with 95.5% precision)

Figure 11(a) shows the performance of each feature set by using the LR classifier:
- content metadata features can only provide preliminary performance
- link properties show the best over-all performance
- “User+Host” shows that combining two feature sets can improve classification results
- “w/o Content” shows that removing content metadata information only slightly degrades the overall performance (1% on the precision rate)

Conclusion

- Spam posting is always in a productive manner as spamming is a professional job nowadays.
- Spam site links in spam posts have much smaller URL length and anchor length due to spammers’ advertising purposes.
- Spam hosts redirected by spam posts often share IP addresses due to the cost-effectiveness considerations by spammers.
- Our spam detection method can achieve precision and recall rates of more than 95% without the help of content-based classifications